Universität Basel
Wirtschaftswissenschaftliche Fakultät

WWZ

November 2016

Self-fulfilling Prophecies in Rank Order Tests

WWZ Working Paper 2016/05

Dragan Ilić

A publication of the Center of Business and Economics (WWZ), University of Basel.

Universität Basel
Peter Merian-Weg 6
4052 Basel, Switzerland
wwz.unibas.ch

Corresponding Author:
Dr. Dragan Ilić
Tel: +41 (0) 61 207 33 57
Mail: dragan.ilic@unibas.ch

Self-fulfilling Prophecies in Rank Order Tests

Dragan Ilić*

November 30, 2016

Abstract

Modeled rank order tests have become a powerful tool to infer discrimination through observational outcome data such as police search success rates or errors in court decisions. The tests predict that outcomes that are born out of prejudice by decisionmakers will violate certain rank order patterns among the treated groups. This paper presents an unnoticed issue in these tests. I advance a rank order model that includes strategic behavior of the treated with respect to the decisionmakers' beliefs about them. This feedback mechanism can give rise to multiple equilibria, which can invalidate the use of the test.

*University of Basel, Faculty of Business and Economics, Peter Merian-Weg 6, 4002 Basel, Switzerland, and ETH Zurich, CER Center of Economic Research, ZUE F11, Zurichbergstrasse 18, 8092 Zurich (e-mail: dragan.ilic@unibas.ch). I am grateful to Georg Nöldeke, George Sheldon, Alois Stutzer, and Brigitte Guggisberg for their help on this project. Funding for this research was generously provided by the WWZ Forum and the National Centre of Competence in Research "On the Move", which is financed by the Swiss National Science Foundation.

1 Introduction

Disentangling the causes for disparate outcomes in observational data is riddled with difficulties. Economics has embraced a host of increasingly sophisticated techniques to answer the question if differences in, say, labor market outcomes between agents like men and women can be attributed, if anything, to taste-based discrimination or statistical discrimination by decisionmakers (Guryan and Charles, 2012). Recently, a promising avenue has been paved by so-called rank order tests. One of the main issues in detecting taste-based discrimination by disparate treatment is that the treated agents are likely to differ in a myriad of unobserved characteristics which can affect the observed outcomes. Rank order tests, a refined specification of outcome tests, circumvent this inference problem by putting the cart before the horse, so to speak. Instead of trying to gather information on the agents' characteristics, outcome tests focus on the agents' outcomes on the notion that the outcomes are an indicator of preceding disparate treatment by malevolent decisionmakers. The idea is that all outcome-relevant information about the agents was processed during the decisionmaking and is thus reflected in the outcome data, obviating the need for collecting microlevel data about the agents.

Consider first a classic outcome test example. Specifically, the decision to grant mortgages, and the question whether that process is racially prejudiced. A bank is to judge applications for mortgages from black and white applicants. During the decision process, that bank constructs some score of applicant creditworthiness which inversely links to default risk. At some level of creditworthiness, the bank will deem that risk acceptable and all mortgage applications that satisfy that level will be granted. Some of the granted mortgages will inevitably default, but at a lower rate than the ones below the acceptance level would have. This mechanism can inform about racial prejudice. If the bank is biased, it holds the disadvantaged group to a higher standard of creditworthiness, driving down the average default rates of their granted applications. One might be tempted to infer prejudice from a comparison between average default rates. After all, if

one group exhibits a lower default rate, the bank obviously foregoes good risks, risks which it "pays" with its taste for discrimination. But because we have not assumed anything about the *distributions* of creditworthiness between the two groups, the average default rates between black and white mortgage owners may differ even if the bank is unbiased (Yinger, 1996). Enter rank order tests. Consider other banks facing the same pool of mortgage applicants. These banks may have different risk preferences; some could be more prudent and others more lenient in granting mortgages. If the banking system as a whole unbiased, however, the racial rank order of the default rates should be the same at all banks. That is to say, if at one bank black applicants have higher default rates than white applicants, black applicants should have higher default rates at any bank.

This example captures the essence of rank order tests for prejudice. We do not need detailed (and hence likely unobtainable) information on each applicant's characteristics; a simple comparison of their outcome data will do. The microeconomic framework that underlies rank order tests was established by Anwar and Fang (2006), who apply their model to motor vehicle search data in Florida. Anwar and Fang check whether the rank order of police officers' search success rates (grouped by officer race) depends on the race of the motorists. If so, their model would imply that officers use suspicion thresholds (for searching a vehicle) that depend on the race of the motorist; evidence of prejudice. This is not the case in their data. The results of the empirical analysis are consistent with the hypothesis of an unprejudiced police force. Addressing the same setting, Close and Mason (2007) apply a similar rank order test to a slightly extended data set and reject the hypothesis of no prejudice. Anbarci and Lee (2014), too, take a close look at prejudice in policing, but are concerned with discretionary behavior in issuing traffic tickets. Their findings imply that officers in Boston are racially biased. Anwar, Bayer, and Hjalmarsson (2010) turn to the question of prejudice in criminal trials. They test their model prediction that in the absence of prejudice, a jury that holds a lower threshold for conviction should convict both black and white defendants on average more than a

jury with a higher threshold. Put differently, the rank order of conviction of defendants of each race should not depend on the type of jury. The authors apply their test to 401 felony trials in Sarasota County in Florida and find evidence of prejudice.[1] Alesina and La Ferrara (2014) collect and analyze a data set on capital punishments in the United States. Their modeling of the courts' decisionmaking process implies that for all racial defendant groups the rank order of error rates (as subsequently uncovered by higher courts) across victim race should be the same if courts are unbiased. Their empirical findings do not satisfy the model prediction, implying the presence of racial prejudice in their data. Finally, Ilić (2016) considers municipalities in Switzerland, each of which decides on their local citizenship applications. Allowing municipalities to vary in their strictness for naturalilzation, unbiasedness implies that the rank order of the applications' rejection rates, grouped by country of origin, should be the same for all municipalities. The empirical analysis exploits an exogenous variation by a court ruling and rejects the hypothesis of no prejudice in a within-municipality-test in one of the six observed municipalities.

The literature on rank order tests shares the implicit assumption that the agents' behavior is not affected by the decisionmakers' beliefs about them. When a decisionmaker interprets an agent's signal to extract information, he combines that signal with a prior, an exogenous belief about that agent's group. This is the statistical discrimination component in these models. In this paper, I show that once we allow for this belief to become endogenous, things can go awry. In their seminal contribution on endogenous beliefs and stereotypes, Coate and Loury (1993) demonstrate how two ex ante identical groups may position in different, Pareto ranked, equilibria. Specifically, negative stereotypes about a group in form of poor beliefs may drive agents to a worse, self-fulfilling equilibrium. I incorporate Coate and Loury's approach by taking the example of Ilić's (2016) rank order model. Municipalities grant citizenships, and applicants make a decision about

[1] The published version of their paper restricts its empirical analysis to exploiting the quasi-random variation in the composition of the seated jury (Anwar, Bayer, and Hjalmarsson, 2012).

an investment to become qualified for naturalization, taking into account the prevailing stereotype about the group they belong to.

2 An Endogenous Rank Order Model

Consider a number of municipal councils $c \in \{c_1, c_{2,...}\}$ that separately evaluate their local immigrants applying for naturalization. Each council faces the same pools of applicants, continuums that are grouped into country of origin $a \in \{a_1, a_2, ...\}$. For the purpose of this paper, consider two municipalities and two applicant groups. The model readily extends to m municipalities and n applicant groups.

Councils want to grant citizenship to applicants only if they are qualified for naturalization. All applicants, qualified or not, prefer to be naturalized, for they would gain the payoff ω. If their application is rejected, their payoff is zero. Councils receive a payoff $x_{q,c,a} > 0$ if they grant a qualified applicant citizenship but are left with a loss of $-x_{u,c,a} < 0$ if they naturalize an unqualified applicant (naturalizations are irreversible). Councils may differ in their payoffs and losses independent of country of origin. This would reflect a variation of strictness or leniency or more lenient in their general tendency to grant citizenship. Councils may also differ in their payoffs dependent on country of origin, reflecting prejudice.

Applicants are not born qualified for naturalization. In order to become qualified, they must invest in some costly skill investment, for example, getting accustomed to local traditions, learning the local language, or reaching some level of human capital. Let the cost of this investment i_a be heterogenous across applicants in each group. The distribution of i_a follows the cumulative distribution function $G_a(\cdot)$. $G_a(\cdot)$ may be group dependent, implying that some groups may have higher investment costs on average because of their socio-economic background. The application itself does not entail costs.

Individual skill investments are imperfectly observable by the councils. Specifically,

councils observe country of origin affiliation a and a noisy signal $\theta \in [0,1]$. This signal is drawn from the interval $[0,1]$ according to the probability density function $f_q(\theta)$ if the applicant is qualified and according to $f_u(\theta)$ if the applicant is unqualified. The signal is assumed to be informative about skill investment in the sense that the distributions $f_q(\cdot)$ and $f_u(\cdot)$ satisfy the Monotone Likelihood Ratio Property (MLRP): $l(\theta) \equiv \frac{f_q(\theta)}{f_u(\theta)}$ is strictly increasing in θ.

The timing of the game is as follows: First, Nature draws the applicants' types, that is, their skill investment cost i_a. After observing their type, applicants make their skill investment decision and then apply for citizenship. Councils then decide whether to grant citizenship based on the applicants' group affilliation and his or her signal of qualification.

Let us first derive the councils' best response. Suppose that a council evaluates an applicant with signal θ from group a, which has an average qualification (according to the council's prior beliefs) of π_a. The posterior probability that such an applicant is qualified follows from Bayes' rule:

$$p(\theta, \pi_a) = \frac{\pi_a f_q(\theta)}{\pi_a f_q(\theta) + (1-\pi_a) f_u(\theta)}$$

The council's expected payoff from granting that applicant citizenship is therefore

$$p(\theta, \pi_a) x_{q,c,a} - [1 - p(\theta, \pi_a)] x_{u,c,a}$$

So citizenship will only be granted if above expression is equal to or greater than zero. Put differently, the council will require a signal threshold value $\overline{\theta}_c(\pi_a)$, the standard required for naturalization, where $\overline{\theta}_c(\pi_a)$ is pinned down by

$$l(\theta) \equiv \frac{f_q(\theta)}{f_u(\theta)} = \frac{1-\pi_a}{\pi_a} \frac{x_{u,c,a}}{x_{q,c,a}} \tag{1}$$

In what follows, we focus on the interior solution of equation (1). How does the threshold value of $\overline{\theta}_c(\pi_a)$ change with the council's prior beliefs? It seems intuitive that

the better the council's prior beliefs, the lower the threshold $\overline{\theta}_c(\pi_a)$ will be set. Following Coate and Loury (1993), the MLRP implies that $\overline{\theta}_c$ is strictly decreasing in π_a:

$$\frac{d\overline{\theta}_c}{d\pi_a} = -l'(\overline{\theta}_c(\pi_a))\frac{x_{u,c,a}}{x_{q,c,a}}\frac{1}{\pi_a^2} < 0 \tag{2}$$

Note that the *level* of the relationship of $\overline{\theta}_c$ on π_a depends on the ratio $\frac{x_{u,c,a}}{x_{q,c,a}}$. If unbiased, this reflects the heterogeneity in strictness among councils.

Let us now turn to the applicants' best response in view of the councils' optimal thresholds $\overline{\theta}_c(\pi_a)$. For an applicant from group a, the skill investment cost amounts to i_a and yields a probability of $1 - F_q(\overline{\theta}_c)$ of receiving citizenship. Investing thus entails an expected payoff of

$$\left[1 - F_q(\overline{\theta}_c)\right]\omega - i_a$$

Without investing, there is also a chance of being (mistakenly) granted citizenship. In that case, the payoff is

$$\left[1 - F_u(\overline{\theta}_c)\right]\omega$$

Investing thus becomes worthwhile if and only if the skill investment costs are lower than the expected benefit subject to the threshold $\overline{\theta}_c$:

$$i_a \leq B(\overline{\theta}_c) \equiv \left[F_u(\overline{\theta}_c) - F_q(\overline{\theta}_c)\right]\omega$$

Note that $B'(\overline{\theta}_c) = \omega\left[f_u(\overline{\theta}_c) - f_q(\overline{\theta}_c)\right] > 0$ only if $\frac{f_q(\theta)}{f_u(\theta)} < 1$. $B(\cdot)$ is therefore a single-peaked function with $B(0) = B(1) = 0$ (see Coate and Loury, p.1225).

We now turn to the equilibrium of this game. Consider first the interaction of one council with one applicant group. Faced with the naturalization threshold $\overline{\theta}$, the fraction of applicants that invests in skills is the fraction of applicants whose investment costs are below $B(\overline{\theta})$, that is

$$G(B(\overline{\theta})) = G(\left[F_u(\overline{\theta}) - F_q(\overline{\theta})\right]\omega)$$

Since $G(B(\cdot))$ is a positive monotone transformation of $B(\cdot)$, it is also single-peaked with $G(B(0)) = G(B(1)) = 0$. An equilibrium exists if the beliefs of the council are self-confirming, that is, when those beliefs induce the applicants to invest precisely at the rate postulated by the beliefs. Formally, an equilibrium is a pair of $(\overline{\theta}^*, \pi^*)$ such that

$$\overline{\theta}^* = \overline{\theta}(\pi^*) \tag{3}$$

and

$$\pi^* = G(B(\overline{\theta}^*))$$

where the first condition describes the equlibrium council behavior and the second condition describes the equilbrium applicant behavior.

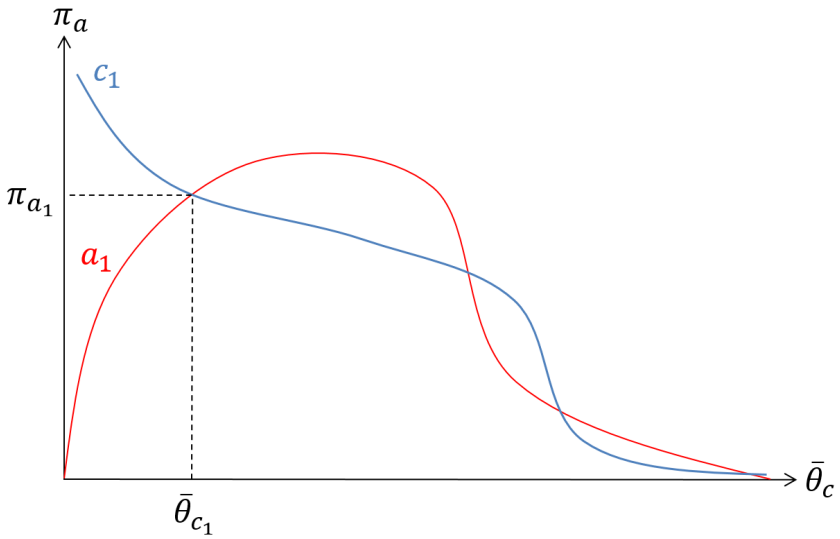

Figure 1: Equilibrium in Applicant and Council Responses

An equilibrium is not necessarily stable in the sense that we can assume an adjustment process of $\pi^{t+1} = G(B(\overline{\theta}^*))$ with $t = 0, 1, 2, ...$(Coate and Loury, p.1126). Local stability is only given if, at the point of intersection, the absolute value of the slope of the applicant best response function is lower than the one of the council's best response function. Also note that the equilibrium - if it exists - can have multiple solutions. This is evident in Figure 1, which sketches the applicant (in red) and council (in blue) response functions $G(B(\overline{\theta}))$ and $[\overline{\theta}(\pi^*)]^{-1}$ (where the latter is the inverse of equation (3) and thus also strictly decreasing in $\overline{\theta}$ because of equation (2)). The multiple solutions involve so-called discriminatory equilibria, equilibria in which a negative belief about a group is self-confirming and associated with a higher standard. In Figure 1, such equilibria can be seen in the downward sloping domain of the applicant best response function. For the remainder of this section, let us focus on the non-discriminatory and stable equlibrium in the upward sloping domain in Figure 1.

Let us now extend the involved parties. First, add a second applicant group a_2, which we define to have higher investment costs on average, that is, $G_{a_2}(i) < G_{a_1}(i)$. As can be seen in Figure 2, this implies that the best response function of group a_2 (in green) lies strictly below the one of group a_1. Equilibrium for group a_2 is thus associated with a higher signal threshold and a lower fraction of qualified applicants. In the upward sloping domain, this inverse relationship always holds: Higher investment costs entail higher signal thresholds and lower shares of qualification.

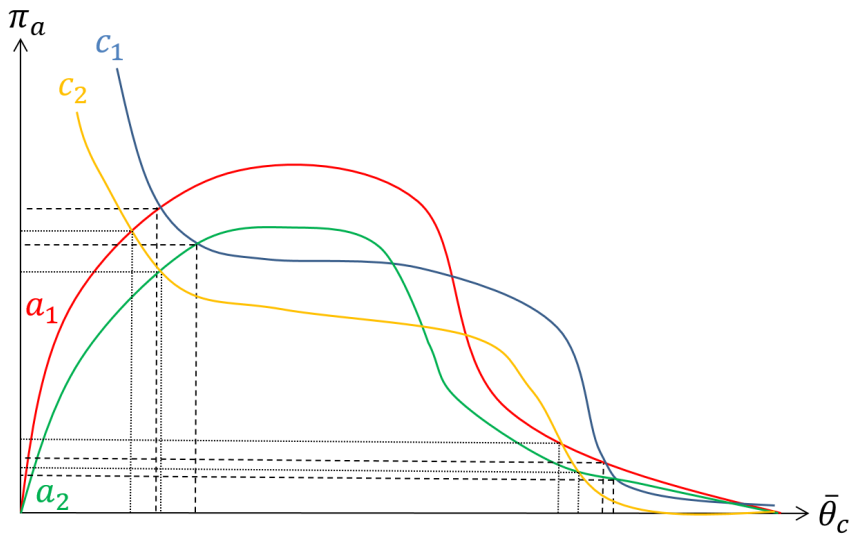

Figure 2: Multiple Equilibria with Two Councils and Two Applicant Groups

In a second step, add a second council c_2. Let this council be more lenient with respect to its requirement for naturalization. From equation (1) we can gather that this may be because council c_2 gains more from naturalizing qualified applicants or because it loses less from mistakenly naturalizing unqualified applicants. In Figure 2, this draws a best response function for council c_2 (in orange) that lies strictly below the one of council c_1. Like in council c_1, in council c_2 the applicant group with the higher investment costs a_2 faces a higher signal threshold and has a lower fraction of qualified applicants than group a_1. The rank order thus remains the same with the introduction of heterogeneity in council strictness. The implied rank order gives rise to a test for prejudice. Applicant groups may differ in investment costs, and councils may differ in strictness. But if councils are not prejudiced, the applicant rejection rates should have the same ranking in all councils. Otherwise, in at least one municipality applicants are not being evaluated by objective criteria. So far, this establishes a conventional rank order test. The reasoning, however, only holds when we consider the upward sloping domain of the applicant best

response function $G(B(\bar{\theta}))$.

3 Multiple Equilibria

The reason why Pareto-worse equilibria may exist in this model is because a council's assessment about the applicants' qualifications is a combination of individual signals and (self-sustaining) beliefs about the applicant groups. If the council holds negative beliefs in a discriminatory equilibrium, the affected applicants are discouraged from investing in requirements for naturalization. Recall that not all equilibria are stable. The discriminatory equlibria are located in the downward sloping domain of the applicant best response function. In that domain, the conditions for stability hold only in equlibria where the council best response function cuts the applicant best response function from above. There is no limit to the number of stable equilibria in this model, but for the purpose of this paper it suffices to look at one stable discriminatory equilibria.

From Figure 2 we can glean that the discriminatory equilibrium implies the same rank order for the qualified fractions of applicants among the two applicant groups as the non-discriminatory equilibrium does. It is easy to show that any stable equilibria will keep the rank orders intact. As long as *all* applicant groups in a given municipality end up in the same ordinal equilibrium, the rank order rationale holds and allows for testing for prejudice among all councils.

Matters are less comforting when, within a council, applicant groups end up in different ordinal equilibria. For a given applicant group, discriminatory equilibria have lower fractions of qualified applicants. In Figure 2 this would mix up the rank orders if only the applicant group with the lower investment cost, a_1, finds itself in the discriminatory equilibrium in either council, c_1 or c_2. The rank order pattern that implies unbiasedness, namely that no matter the council, the share of qualified applicants is higher among a_1 than among a_2, fails to hold even though no council is prejudiced. Testing for prejudice

with the rank order test would mistakenly detect taste-based discrimination when, in fact, negative beliefs cause the breach in rank order.

4 Conclusion

This paper reveals a neglected issue in rank order tests for prejudice. Once we allow for the decisionmakers' beliefs to affect the agents' behavior, multiple equilibria can arise. The rank order test cannot differentiate between effects of prejudice and effects of applicants ending up in different equilibria; the two are observationally equivalent. This conflation is not restricted to the particular model presented in this paper. Any setting where the decisionmakers' beliefs about the agents' quality affect their incentives to acquire said quality is prone to exhibit multiple equilibria.

References

Alesina, Alberto, and Eliana La Ferrara. 2014. "A Test of Racial Bias in Capital Sentencing." *American Economic Review*, 104(11): 3397–3433.

Anbarci, Nejat, and Jungmin Lee. 2014. "Detecting Racial Bias in Speed Discounting: Evidence from Speeding Tickets in Boston." *International Review of Law and Economics*, 38: 11–24.

Anwar, Shamena, and Hanming Fang. 2006. "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence." *American Economic Review*, 96(96): 127–151.

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson. 2010. "Jury Discrimination in Criminal Trials." *Queen Mary University, School of Economics and Finance, Working Paper No. 671.*

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson. 2012. "The Impact of Jury Race in Criminal Trials." *The Quarterly Journal of Economics*, 127(2): 1017–1055.

Close, Billy R., and Patrick L. Mason. 2007. "Searching for Efficient Enforcement: Officer Characteristics and Racially Biased Policing." *Review of Law and Economics*, 3(2): 263–321.

Coate, Stephen, and Glenn C. Loury. 1993. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review*, 83(5): 1220–1240.

Guryan, Jonathan, and Kerwin Kofi Charles. 2013. "Taste-based or Statistical Discrimination: The Economics of Discriminatino Returns to Its Roots." *The Economic Journal*, 123: F417–F432.

Ilić, Dragan. 2016. "Prejudice in Naturalization Decisions: Theory and Evidence." WWZ Working Paper 2016/04.

Yinger, John. 1996. "Why Default Rates Cannot Shed Light on Mortgage Discrimination." *Cityscape: A Journal on Policy Development and Research*, 2(1): 25–31.